High Performance Data Transfer and Monitoring for RHIC and USATLAS

Modern nuclear and high energy physics experiments yield large amounts of data and thus require efficient and high capacity storage and transfer. BNL, the hosting site for RHIC experiments and the US ATLAS Tier 1 center, plays a pivotal role in transferring to and from other sites in the US and around the world for data distribution and processing. Each component in the infrastructure from data acquisition system to local analysis facility must be monitored, tested, and tuned to transfer such a sheer volume of data often over long distances. Ultimate performance can be reached by performing hardware optimization, TCP tuning, transfer application tuning and network architecture adjustments

Authors:
Packard, Jay (BNL)
Yu, Dantong (BNL)
Katramatos, Dimitrios (BNL)
Lauret, Jerome (BNL)
Shroff, Kunal (BNL)

Coauthors:
Purschke, Martin (BNL)
Watanabe, Yasushi (Riken)
Woo, Joon (Kisti)
Kim, Donyun (Kisti)
Betts, Wayne (BNL)
DeStefano, John (BNL)
Hover, John (BNL)
McKee, Shawn (University of Michigan)

Apply optimal network tunings based on source to destination distance, bandwidth requirements, and previous test results.

If residual performance exists, collaborative work with network engineers at both ends is essential to troubleshoot performance issues such as packet drop or inefficient or non symmetric routing. May need to:
- Understand network topology in depth including BW pipe sizes, MTU, and DSCP.
- Understand hosts in depth including network parameters, firewall configuration, disk IO, NIC make and model.
- May need to upgrade, reconfigure, or reroute network devices.
- May need to bypass firewall if severe performance degradation.

IMPORTANT: Allow acks in and syn-acks out in iptables

Need to test multiple layers (using iperf, gridftp, perfsonar):
- UDP to obtain packet loss statistics
- memory to memory (TCP) to determine network performance
- Memory to disk to determine destination disk performance
- Disk to memory to determine source disk performance
- Disk to disk to determine likely production performance

Need to test parameter combinations (using BNL's MetaPerf script):
- Number of streams
- Kernel TCP parameters
- NIC parameters

Need to test at multiple hops along the way:
- Host inside firewall
- Host outside firewall
- Intermediate WAN hosts if bottleneck is discovered

Tune → **Test** → **Monitor** → **Troubleshoot** → **Results**

Before and after results from tuning are usually drastic, often yielding more than ten fold increase in transfer rate. This was seen in all of our tests:
1) From BNL to and from it's USATLAS tier 2 sites.
2) From BNL RHIC experiment data acquisition systems to the computing center in Japan (CCJ) on behalf of the PHENIX experiment.
3) From BNL RHIC experiment data acquisition systems to the Korea Institute of Science and Technology Information (KISTI) on behalf of the STAR Experiment.

The outcomes of this work are being integrated into:
1) The ATLAS production and analysis framework to allow USATLAS regional centers to do data processing.
2) The global data taking and reconstruction framework for RHIC experiments to leverage the computing resources of RHIC international collaborators.
3) The potential of establishing RHIC Tier1 centers and data redistribution hubs In Asia (STAR already has 27% of its institutional workforce located in Asia).

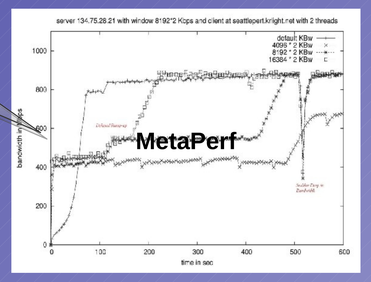
Identify bottlenecks.

Final Throughput Test Results

- BNL -> USATLAS Tier2s: 3.2 Gb/s disk to disk
- BNL -> CCJ: 1.3 Gb/s disk to disk
- BNL -> KISTI: 1 Gb/s disk to disk

The fastest disk to disk rates ever acheived between the US and Asia!

Study the effect of bandwidth over time due to changes in network parameters.



BNL

Test from within firewall and from outside firewall to determine firewall degredation.

Test to intermediate host to isolate problems between portions of the WAN.

ESNet / Internet2

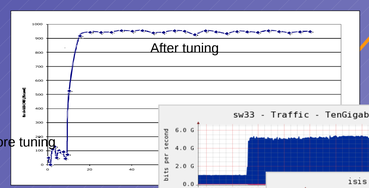
SINET

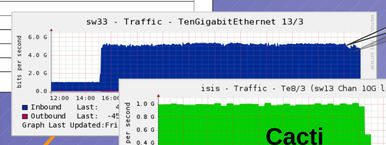
PNW Gigapop

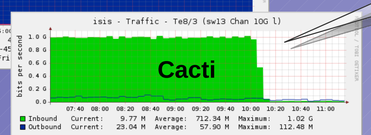
GLORIAD

4 Gb/s to HPSS

1 Gb/s to KISTI

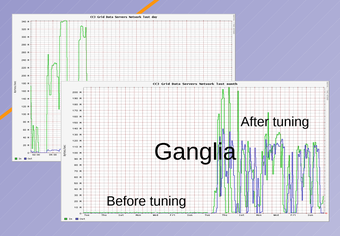


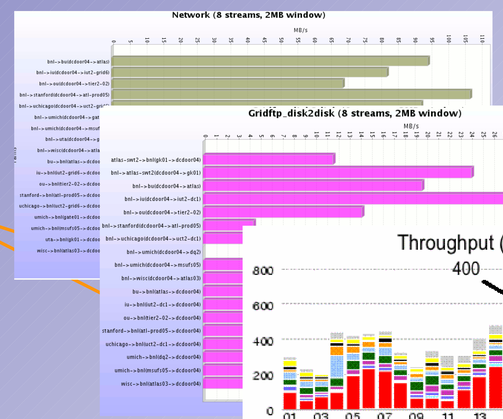




KISTI at Korea for STAR

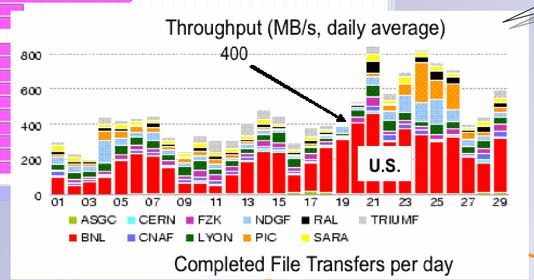
CCJ at Japan for PHENIX





Individual memory and disk-to-disk monitoring.

Aggregate disk-to-disk monitoring.



USATLAS Tier 2 site

USATLAS Tier 2 site...